CHALMERS
UNIVERSITY OF TECHNOLOGY

Predicting permeability via statistical learning on higher-order microstructural information

SCIENTIFIC REPORTS
nature research

Check for updates

OPEN

Predicting permeability via statistical learning on higher-order microstructural information

Magnus Röding[1,2]✉, Zheng Ma[3] & Salvatore Torquato[4]

Quantitative structure–property relationships are crucial for the understanding and prediction of the physical properties of complex materials. For fluid flow in porous materials, characterizing the geometry of the pore microstructure facilitates prediction of permeability, a key property that has been extensively studied in material science, geophysics and chemical engineering. In this work, we study the predictability of different structural descriptors via both linear regressions and neural networks. A large data set of 30,000 virtual, porous microstructures of different types, including both granular and continuous solid phases, is created for this end. We compute permeabilities of these structures using the lattice Boltzmann method, and characterize the pore space geometry using one-point correlation functions (porosity, specific surface), two-point surface-surface, surface-void, and void-void correlation functions, as well as the geodesic tortuosity as an implicit descriptor. Then, we study the prediction of the permeability using different combinations of these descriptors. We obtain significant improvements of performance when compared to a Kozeny-Carman regression with only lowest-order descriptors (porosity and specific surface). We find that combining all three two-point correlation functions and tortuosity provides the best prediction of permeability, with the void-void correlation function being the most informative individual descriptor. Moreover, the combination of porosity, specific surface, and geodesic tortuosity provides very good predictive performance. This shows that higher-order correlation functions are extremely useful for forming a general model for predicting physical properties of complex materials. Additionally, our results suggest that artificial neural networks are superior to the more conventional regression methods for establishing quantitative structure–property relationships. We make the data and code used publicly available to facilitate further development of permeability prediction methods.

The study of how the microstructural morphology of random, heterogeneous, porous materials affects their effective properties, i.e., determining quantitative structure–property relationships, is key for the understanding and prediction of the physical properties of complex materials[1]. Specifically, understanding how fluid transport properties are related to the microstructure of a porous medium is crucial in a wide range of areas e.g. geological events[2], polymeric composites for packaging materials[3], catalysis, filtration and separation[4], energy, fuels, and electrochemistry[5], fiber and textile materials for health care and hygiene[6], and porous, biodegradable polymer films for controlled release of medical compounds[7]. Numerous efforts in determining the physical properties of complex materials have been made since the early work of Maxwell[1,8–11], and such investigations have been enhanced due to the availability of high-resolution 3D images of various types of materials microstructures using X-ray nanotomography[12,13] or focused ion beam scanning electron microscopy[14], and nuclear magnetic resonance as well[15,16].

[1]RISE Research Institutes of Sweden, 41276 Göteborg, Sweden. [2]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 41296 Göteborg, Sweden. [3]Department of Physics, Princeton University, Princeton, NJ 08544, USA. [4]Department of Chemistry, Department of Physics, Princeton Institute for the Science and Technology of Materials, and Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA. ✉email: magnus.roding@ri.se

The porosity ϕ (volume fraction of the pore phase) and specific surface s (pore-solid interface area per unit volume) are perhaps the most basic geometrical characteristics. These two characteristics are the most frequently used in empirical expressions for the permeability. The Kozeny-Carman equation[17,18] is the most notable example, usually written as

$$k = \frac{\phi^3}{cs^2}, \tag{1}$$

where k is the permeability and c is the Kozeny-Carman constant. However, the remarkably simple form comes with great limitations. The Kozeny-Carman constant was found not to be a universal quantity. It does not only vary for different systems, but can also depend on the porosity[19]. Additionally, it does not distinguish portions of pore space that carries significant flow from portions that do not[1].

To tackle this difficulty, countless modified versions of the original Kozeny-Carman equation have been proposed. However, these models are usually *ad hoc* and only applicable to a specific class of structures[20]. More importantly, although in many cases tortuosity is incorporated in the Kozeny-Carman constant[21–24], it usually only depends on the porosity alone in simplified models, thus the final expression of the permeability is essentially nothing more than a function of porosity and specific surface, i.e., $f(\phi)/s^2$. However, it is well-known that the microstructure is highly degenerate given only porosity and specific surface[25,26], which leads to a wide range of permeabilities as we show later. Thus, any function of the form $f(\phi)/s^2$ cannot be a general predictor and suffers from the intrinsic variances in the set of infinite degenerate microstructures.

Indeed, accurate prediction of the effective physical properties of the porous media requires a complete quantitative characterization of the microstructure in d-dimensional Euclidean space $\mathbb{R}^d$ via a variety of n-point correlation functions[1]. However, while such complete structural information about the medium is generally not available, reduced information in the form of lower-order correlation functions is often very beneficial. Two-point void-void and three-point void-void-void correlation functions have been used to produce both bounds and estimates for the effective electrical conductivity, diffusion coefficient and permeability[27–36]. In addition to the void-void correlation function, two-point surface-surface and surface-void correlation functions (where the surface is the interfacial surface between two phases) can also be defined and provide improved reconstructions of two-phase media from imaging data[37,38], as well as sharper bounds on permeability compared to only using the void-void correlation function[1].

On the other hand, it has been shown that permeability can be simply connected to the electrical formation factor of the porous material[39,40]. This has been proved rigorously by Avellaneda and Torquato[41]. However, from a prediction point of view, the formation factor itself needs to be measured experimentally or solved numerically, thus is not that helpful for establishing an explicit link to the microstructure. Although the formation factor is related to the hydraulic tortuosity[42], the later also requires heavy computations.

As a complement to rigorous approaches to estimate effective properties from the microstructure, data-driven methodologies to establish structure–property relationships are increasingly being used[43–49]. The rapid increase in computational resources facilitates the computation of effective properties for very large data sets (hundreds or thousands) of different microstructures. Moreover, as noted above, affordable high-resolution 3D digitized images of actual microstructures provide valuable data sets. As a consequence, it becomes manageable to generate large numbers of realistic virtual microstructures, and using those to perform exploratory computational screening of structure–property relationships. For example, van der Linden et al.[43] use a data set of 536 virtual granular materials, compute 27 geometrical descriptors and use log-linear regression and other statistical learning methods as well as different variable selection schemes to understand the usefulness of the different descriptors for predicting permeability in these systems. Stenzel et al.[44] study effective conductivity prediction in 43 virtual realizations of a stochastic spatial network model structure, using porosity and different tortuosity and constrictivity measures. This study was extended to 8,119 microstructures[50], which is likely the largest study published before, and the same data set was used again later to predict effective conductivity and permeability[45]. Barman et al.[46] studied effective diffusivity prediction in 36 virtual porous polymer films using tortuosity and constrictivity. In a different direction, there are several attempts to use 2D and 3D convolutional neural networks (CNNs) to extract information directly from the binary image data describing the structure[47–49,51–53] in order to predict effective properties. However, these models are usually difficult to interpret and hard to rescale.

In this work, we are primarily interested in the predictive power of the information content contained in different microstructural descriptors. Specifically, we investigate the two-point surface-surface, surface-void, and void-void correlation functions, and also porosity, specific surface, and geodesic tortuosity using different regression methods. Unlike the hydraulic tortuosity mentioned above, the geodesic tortuosity is a purely geometric quantity that can be computed efficiently, and has been shown to be superior for diffusivity prediction[46]. We compare different regression methods, including conventional linear regression with linear and quadratic terms, as well as deep artificial neural networks (deep learning). While conventional linear regression has an advantage in so far as the transparency of the prediction mechanism, deep learning has the potential to extract nearly the full information content of the descriptors, providing insight into the utility of the different descriptors for establishing the structure–property relationship. We find that the information content contained in these two-point correlation functions and geodesic tortuosity are indeed helpful to overcome the difficulty of applying a unique Kozeny-Carman-type equation to a variety of distinct microstructures, by yielding much better prediction performance. Moreover, our results suggest that artificial neural networks are superior to the more conventional regression methods for establishing quantitative structure–property relationships.

Consistent with the purpose of the paper, we have generated a large data set of virtual, porous, isotropic, and stationary microstructures of three different types, based on (i) thresholded Gaussian random fields, (ii) thresholded spinodal decomposition simulations of phase separation, and (iii) non-overlapping ellipsoid systems.

Varying porosity and length scale and other parameters, we generate 10,000 structures of each of the three types, yielding a data set of 30,000 virtual microstructures in total. This is likely to be the largest data set of virtual microstructures ever created for studying permeability prediction, and covers both granular (ellipsoids) and continuous solid phases to provide a broad variability in the pore space geometry. Fluid flow is simulated using the lattice Boltzmann method. The large number of simulated microstructures makes it feasible to use not only scalar descriptors but also high-dimensional descriptors such as the two-point correlation functions, while still avoiding the well-known 'curse of dimensionality' in regression caused by having too many dimensions but too little data. To facilitate further investigation and development of permeability prediction methods, we make the microstructural descriptors, the computed permeabilities, the trained models, and the code used herein publicly available[54].

The paper is organized as follows. First, we introduce necessary definitions for the geometric descriptors used throughout the paper. Second, we describe how the virtual microstructures are generated, and the flow simulations and computations of permeability are described. Third, computation of the different microstructural descriptors is covered. Fourth, the prediction models for permeability are investigated. Finally, we make concluding remarks and discussions.

Background and definitions

Geodesic tortuosity. We compute geodesic tortuosity in the flow direction according to Barman et al.[46] in the following manner. As a first step, a pointwise geodesic tortuosity is computed as $\tau(\mathbf{x}) = d(\mathbf{x})/d$. Here, d is the length of the microstructure in the flow direction, and $d(\mathbf{x})$ is the length of the shortest path from any inlet pore to any outlet pore through $\mathbf{x}$. The shortest path is calculated as the sum of two geodesic distance transforms computed in the pore space of the binary voxel array: one using the set of edge voxels constituting the inlet pores as seeding points, and the other using the set of edge voxels constituting the outlet pores as seeding points. Let $\mathbb{P}$ be the set of voxels for which both geodesic distances are finite, i.e., the set of pore voxels connected to both inlet and outlet. Then, the geodesic tortuosity τ can be computed as

$$\tau = \left(\frac{1}{|\mathbb{P}|} \int_{\mathbf{x}\in\mathbb{P}} \frac{d\mathbf{x}}{\tau^2(\mathbf{x})} \right)^{-1/2}. \tag{2}$$

In Barman et al.[46], it was found that accounting for both inlet and outlet in this manner is superior (in terms of diffusivity prediction) to just accounting for the inlet as is commonly done[44,55]. Tortuosity calculations were implemented in Matlab (Mathworks, Natick, MA, US).

Correlation functions. Let $\mathcal{I}(\mathbf{x})$ be the indicator function for the void phase (pore space) $\mathcal{V}_1$, i.e.,

$$\mathcal{I}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathcal{V}_1, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The two-point void-void correlation function is then generally defined by

$$S_2(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathcal{I}(\mathbf{x}_1)\mathcal{I}(\mathbf{x}_2) \rangle. \tag{4}$$

For statistically homogeneous materials, S_2 is only dependent on the vector difference $\mathbf{r} = \mathbf{x}_2 - \mathbf{x}_1$. Further, if the material is also statistically isotropic, S_2 is only dependent on the radial distance $r = |\mathbf{r}|$. Introducing a notation that is consistent with the other correlation functions defined below, the two-point void-void correlation function is now defined as

$$F_{\mathrm{vv}}(r) = \langle \mathcal{I}(\mathbf{x})\mathcal{I}(\mathbf{x}+\mathbf{r}) \rangle, \tag{5}$$

where the average is taken over all $\mathbf{x}$ and over all $\mathbf{r}$ with magnitude r. We proceed to the correlation functions involving the interfacial surface. Let $\mathcal{M}(\mathbf{x})$ be the interface indicator function defined by[1]

$$\mathcal{M}(\mathbf{x}) = |\nabla \mathcal{I}(\mathbf{x})|. \tag{6}$$

Still assuming ergodicity and statistical isotropy, the surface-void correlation function can be written as

$$F_{\mathrm{sv}}(r) = \langle \mathcal{M}(\mathbf{x})\mathcal{I}(\mathbf{x}+\mathbf{r}) \rangle, \tag{7}$$

and the surface-surface correlation function can be written as

$$F_{\mathrm{ss}}(r) = \langle \mathcal{M}(\mathbf{x})\mathcal{M}(\mathbf{x}+\mathbf{r}) \rangle. \tag{8}$$

Importantly, the information content of one-point correlation functions (porosity ϕ and specific surface s) is automatically encoded into these two-point correlation functions. When r goes to infinity, $F_{\mathrm{vv}}(r)$, $F_{\mathrm{sv}}(r)$ and $F_{\mathrm{ss}}(r)$ converge to ϕ^2, $s\phi$ and s^2 respectively. Interestingly, the slope of $F_{\mathrm{sv}}(r)$ at the origin is proportional to the integrated mean curvature of the system[56], which has recently been shown to be a useful predictor of both permeabilities[57] and diffusion coefficients[58].

Accurate and robust computation of F_{vv}, F_{sv}, and F_{ss} from discretized structures is a non-trivial task (in particular for the later two). The calculations recently became accessible due to the algorithms devised by Ma and Torquato[56]. There, the calculations involving the interfacial surface are performed using a scalar field which when thresholded yields the corresponding two-phase medium. The details of the algorithms and the software can be found in Ref[56].

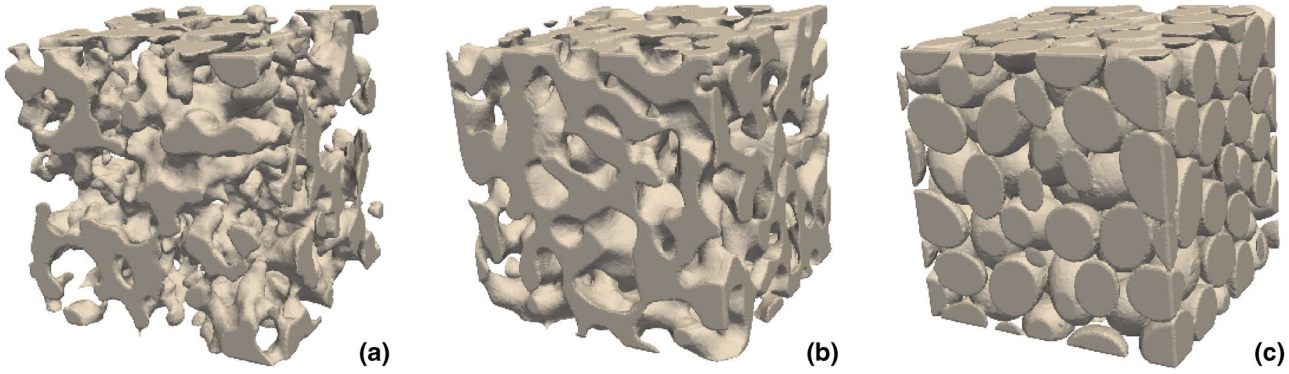

Figure 1. Examples of structures, showing (**a**) Gaussian random field structure with $\phi = 0.7$, (**b**) spinodal decomposition structure with $\phi = 0.5$, and (**c**) non-overlapping ellipsoid structure with $\phi = 0.3$. The figure is produced using ParaView 5.4.1 (http://www.paraview.org, freely available without permission).

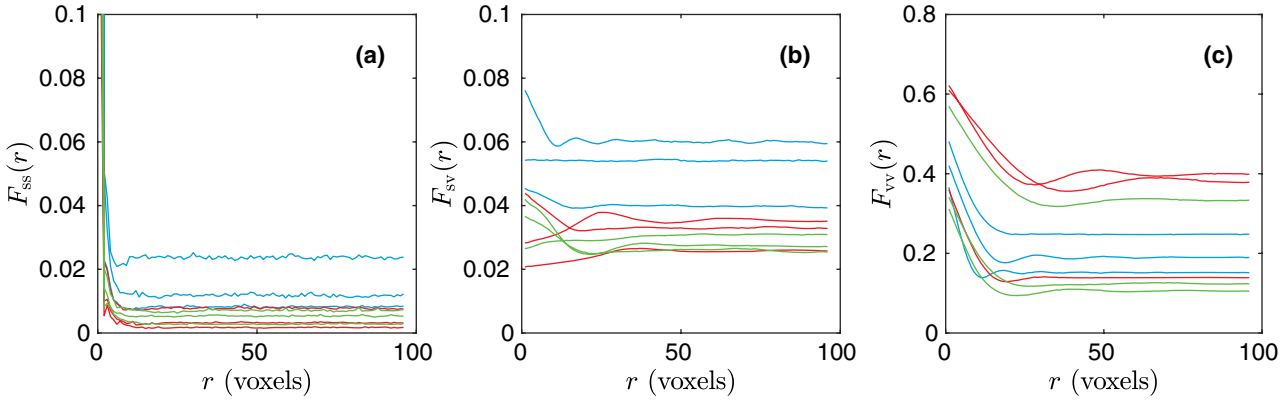

Figure 2. Some examples of (**a**) F_{ss}, (**b**) F_{sv} , and (**c**) F_{vv} correlation functions. The examples are taken from the three different types of generated microstructures, i.e., thresholded Gaussian random fields (blue), thresholded spinodal decomposition simulations of phase separation (red), and non-overlapping ellipsoid systems (green).

Microstructure data preparation

Microstructure generation. To achieve a large, representative data set, three different types of microstructures that are commonly studied in the materials literature are generated, including (i) thresholded Gaussian random fields, (ii) thresholded spinodal decomposition simulations of phase separation, and (iii) non-overlapping (hard) ellipsoid systems. We simulate 10,000 realizations for each type, with porosities ϕ selected uniformly in $0.3 \leq \phi \leq 0.7$ and varying characteristic length scales. In the end, all structures are converted to N^3 binary voxel arrays with $N = 192$ voxels. In Fig. 1, one example of each type of structure is shown. We verified that the choice of the system volume size is both computationally manageable and representative. The correlation functions are evaluated for integer radii value bins r from 1 to 96 voxels. In Fig. 2, some examples of correlation functions are shown. Note that these correlation functions are considerably distinct from each other, as seen by their different magnitudes and functional shapes. On the other hand, they have already converged to the large-r limits within the sample size. The details of how these samples are generated are presented in the following subsections.

Gaussian random fields. Gaussian random fields are generated according to Lang and Potthoff[59]. Assuming that we wish to simulate a Gaussian random field $\mathcal{G}(\mathbf{x}), \mathbf{x} \in \mathbb{R}^3$, with mean zero and covariance function $\Psi(\mathbf{x}, \mathbf{y})$, it utilizes the fact that the covariance function can be written

$$\Psi(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^3} e^{-2\pi i \langle \mathbf{p}, \mathbf{x}-\mathbf{y} \rangle} \gamma(\mathbf{p}) d\mathbf{p}, \tag{9}$$

where $\gamma(\mathbf{p})$ is the spectral density of the Gaussian random field and $\langle \cdot, \cdot \rangle$ is the inner product. We wish to generate structures with length scale parameter L and resolution N^3 voxels. Letting $\delta = L/N$ and letting FFT and FFT^{-1} denote the forward and inverse 3-dimensional Fast Fourier Transforms, this can be performed in the following fashion: Generate an array W where all elements are independent and normal distributed with mean zero and standard deviation δ^{-3} (white noise). Compute FFT(W). Define the Fourier space grid by $\mathbf{p} = (p_1, p_2, p_3)$, where $p_1 \in \{-N/(2L), (-N/2+1)/L, ..., (N/2-2)/L, (N/2-1)/L\}$ and likewise for p_2 and p_3. Compute $\gamma(\mathbf{p})$ on the

grid. Compute $U = \mathrm{FFT}(\mathrm{W})(\mathbf{p}) \times \gamma(\mathbf{p})^{1/2}/\mathrm{L}^3$. Then, obtain the Gaussian random field as $\mathrm{FFT}^{-1}(U)$. We use several different spectral densities. For type (I),

$$\gamma(\mathbf{p}) = \left[1 + \left(p_1^2 + p_2^2 + p_3^2\right)^l\right]^{-n} \tag{10}$$

for $n = 1.95$ and $l = 1.85$ (power-law)[59,60]. For type (II),

$$\gamma(\mathbf{p}) = \exp\left[-\alpha^2\left(p_1^2 + p_2^2 + p_3^2\right)^{1/2}\right], \tag{11}$$

for $\alpha = 1.75$ (exponential). For type (III),

$$\gamma(\mathbf{p}) = \exp\left[-\alpha^2\left(p_1^2 + p_2^2 + p_3^2\right)\right], \tag{12}$$

for $\alpha = 1.25$ (Gaussian). For type (IV),

$$\gamma(\mathbf{p}) = \begin{cases} 1, & \text{if } p_1^2 + p_2^2 + p_3^2 \leq \rho^2, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where $\rho = 1.25$ (circular top-hat). The parameters are chosen such that the corresponding Gaussian random fields have approximately the same characteristic spatial scale. For each spectral density, 2,500 structures are generated using uniformly distributed values of L, $4 \leq L \leq 16$. Thresholding then converts the scalar fields to corresponding two-phase media. To obtain a microstructure with a prescribed porosity ϕ, the threshold is chosen to be an appropriate percentile of the values of $\mathcal{G}$. The method is implemented in Matlab (Mathworks, Natick, MA, US). The execution time is approximately 1 s (single core) for each structure.

Spinodal decomposition. The lattice Boltzmann method[61,62], a numerical framework for solving partial differential equations based on kinetic theory, is used to simulate phase separation kinetics (spinodal decomposition) using the Navier–Stokes and Cahn–Hilliard equations. Very briefly, the time evolution of a spatially dependent concentration $C(\mathbf{x}, t)$, $0 \leq C \leq 1$, is described by

$$\frac{\partial C}{\partial t} + \mathbf{u} \cdot \nabla C = M \nabla^2 \mu. \tag{14}$$

As an initial condition, the values of C are uniformly distributed in $0 \leq C \leq 1$, independently in all grid points. The phase separation is the coarsening of regions with $C \approx 0$ and $C \approx 1$ (ideally equal to 0 and 1). Here, $\mathbf{u}$ is a fluid velocity governed by the Navier–Stokes equations, M is a mobility, i.e., a diffusion coefficient, and μ is the chemical potential. The simulation is performed using a dimensionless time step unity and both the density ratio and viscosity ratio between the phases are unity. The interface width, i.e., the characteristic length scale of the transition between the phases is 5 voxels. The simulations are performed in the resolution 96^3 voxels with periodic boundary conditions, and are run until an appropriate degree of coarsening is obtained. The number of iterations K is chosen randomly between 5 and 20,000 such that $K^{1/3}$ is approximately uniformly distributed; this is because according to the Lifschitz-Slyozov law, the typical length scale in the structure will be proportional to the cubic root of the simulation time. After terminating the simulation, the solutions are upscaled to 192^3 voxels and thresholded to obtain the desired porosity. The spinodal decomposition simulations are run using in-house software[61,62] with efficient scaling to many cores using the MPI interface. The average execution time is approximately 13 min (32 cores) for each structure, and up to 60 min for the longest computation.

Non-overlapping ellipsoids. Random configurations of non-overlapping, hard ellipsoids are generated using a hard particle Markov Chain Monte Carlo (MCMC) algorithm. The Perram-Wertheim criterion[63] for two ellipsoids of arbitrary orientation is used for overlap detection. First, particles are assigned uniformly distributed locations and orientations (the latter encoded using a quaternion representation). Second, the configurations are relaxed by sequentially performing random translations of all particles and then random rotations of all particles until no two particles overlap. Proposed translations and rotations are only accepted if they lead to a lower or equal degree of overlap for the considered particle. These "local" stochastic optimization steps eventually lead to a "global" optimization resulting in no overlap. Third, the configurations are equilibrated by performing a large number of random translations and rotations, ensuring a distribution in location and orientation that is as uniform as possible. Now, if the desired porosity ϕ is larger than 0.50, non-overlapping configurations can be generated easily at constant porosity as described above. Otherwise, as a final step, the configuration is further compressed in small steps, $\Delta\phi = 10^{-5}$, until the target porosity ϕ_{target} is reached (in some cases, the configuration becomes jammed before reaching ϕ_{target}). The proposed translations are normal distributed with standard deviation σ_{t} in each direction. The proposed rotations are normal distributed with standard deviation σ_{r} in a random direction. In every step, σ_{t} and σ_{r} are chosen in an adaptive fashion to aim for an acceptance probability of 0.25. The number of ellipsoids M is distributed in $8 \leq M \leq 512$ such that $M^{1/3}$ is approximately uniformly distributed, yielding an approximately uniform distribution of length scales. Further, the ellipsoids have semi-axes $(1, 1, \eta)$ where η is uniform in $0.25 \leq \eta \leq 1$ (oblate) with probability 0.5 and otherwise uniform in $1 \leq \eta \leq 4$ (prolate). The random microstructures are generated using in-house developed software implemented in Julia (http://www.julialang.org)[64] and available in a Github repository (https://github.com/roding/whitefish_generation, version 0.2). The average execution time is approximately 1 min (single core) for each structure, and up to 30 min for the longest computation. The obtained configurations are further smoothed with a Gaussian filter

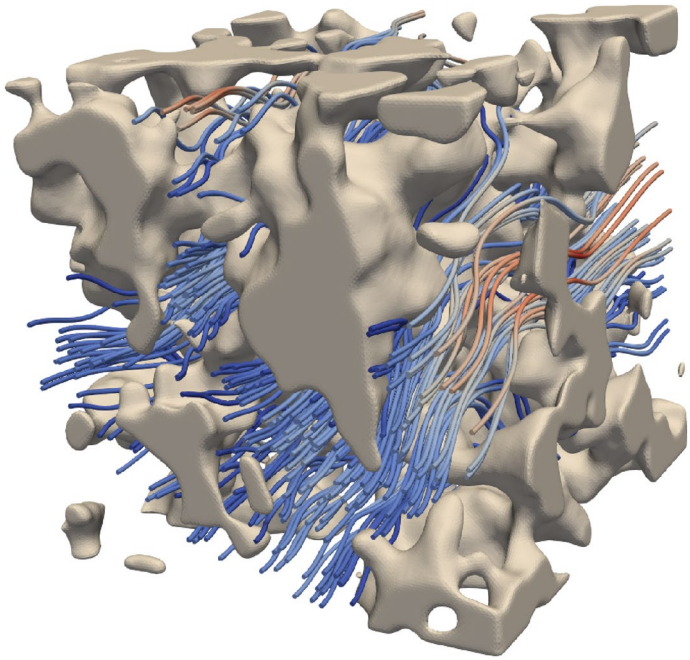

Figure 3. An example of a simulated steady state flow through a Gaussian random field microstructure with porosity $\phi = 0.7$. Regions with slow and fast flow are indicated by blue and red flow lines, respectively. The figure is produced using ParaView 5.4.1 (http://www.paraview.org, freely available without permission).

with $\sigma = 3$ voxels and thresholded again to regain the original porosity; the reason for this is that computation of some of the correlation functions requires the binary structures to be described as a thresholded version of smooth scalar fields.

Flow simulations. The lattice Boltzmann method[61,62], a numerical framework for solving partial differential equations based on kinetic theory, is used to simulate fluid flow through the structures. The Navier–Stokes equations for pressure-driven flow are solved for the steady state using no-slip, bounce-back boundary conditions on the solid/liquid interface and periodic boundary conditions orthogonal to the flow direction. We use the two relaxation time collision model with the free parameter $\lambda_{eo} = \frac{3}{16}$, which guarantees that the computed permeability is independent of the relaxation time (and thus the viscosity)[65]. The relaxation time $\tau = -\frac{1}{\lambda_e}$ is kept at 1.25. The flow is driven by constant pressure difference boundary conditions across the structure in the primary flow direction[66], and a linear gradient is used as initial condition. The computational grid coincides with the voxels of the binary structure, i.e., it has 192^3 grid points. After convergence to steady state flow, the permeability k is obtained from Darcy's law,

$$\bar{u} = -\frac{k\Delta p}{\mu d}. \tag{15}$$

Here, $\bar{u}$ is the average velocity, Δp is the applied pressure difference, μ is the dynamic viscosity, and d is the length of the microstructure in the flow direction. The permeability is independent of the fluid and the pressure difference and a property solely of the microstructure provided that the Reynolds number is sufficiently small ($\mathrm{Re} < 0.01$), which also ensures that the velocity is proportional to the pressure difference. The computed permeabilities have units voxels^2, where the voxels have unit length.

Convergence of the computation is assessed in the following fashion. The energy of the fluid flow field, i.e. the integral of the squared velocity magnitude in the pore space, is computed for each iteration. In each iteration, the coefficient of variation (the standard deviation divided by the mean) of this energy is computed for the latest 500 iterations. The computation is terminated once this coefficient of variation reaches below 10^{-4}. The convergence criterion is the same throughout, although the number of iterations required for convergence differs between different types of microstructures. The mean number of iterations needed are approximately 2,840 for Gaussian random fields, 2,330 for spinodal decompositions, and 2,270 for non-overlapping ellipsoids. However, the number of iterations is also highly dependent on e.g. porosity and varies approximately in the range 1,000 to 10,000 for all types of microstructures. The average execution time is 97 s (utilizing 32 cores, with efficient scaling using the MPI interface).

Figure 3 illustrates the result of a flow simulation in one of the Gaussian random field structures.

We choose 4,500 miscrostructures, 1,500 for each type, and plot their scaled permeabilities ks^2 versus porosities ϕ in Fig. 4. It is noteworthy that although a clear overall trend can be seen, the scaled permeability is never a function of ϕ alone. In fact, we observe that for the same porosity, the largest scaled permeability is approximately

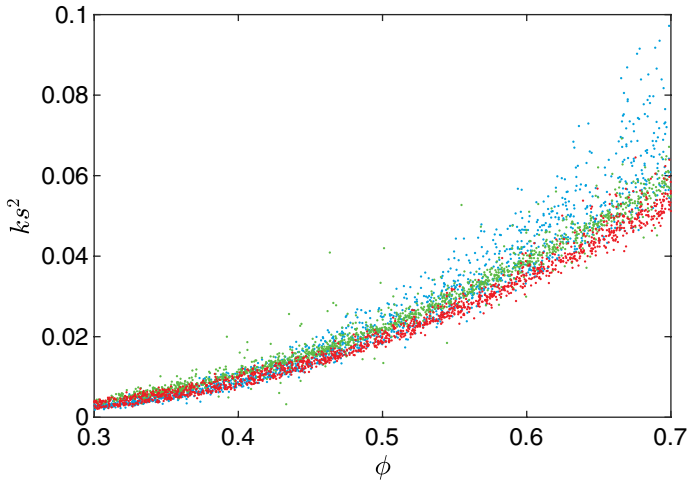

Figure 4. A scatter plot of scaled permeabilities ks^2 versus porosities ϕ for 4,500 microstructures, 1,500 for each type, i.e., thresholded Gaussian random fields (blue), non-overlapping ellipsoid systems (green) and spinodal decomposition simulations (red).

No.	Descriptors	Input size
1	ϕ, s, τ	3
2	F_{ss}	96
3	F_{sv}	96
4	F_{vv}	96
5	F	96
6	F_{ss}, F_{sv}, F_{vv}	288
7	$F_{ss}, F_{sv}, F_{vv}, \tau$	289

Table 1. Descriptors for prediction of the permeability k.

twice as large as the smallest one. This degeneracy of microstructures clearly show why the Kozeny-Carman equation and some of its modifications typically fail for general structures. Consequentially, more detailed information is needed to pinpoint the true permeability on this "band".

Another interesting observation is that the scaled permeabilities of spinodal decomposition patterns are almost always lower than those of the other two types. It has been shown that spinodal decomposition gives rise to hyperuniform structures in the scaling region[67]. This observation is consistent with the fact that hyperuniform structures cannot tolerate large "holes" and the pore space is more evenly distributed compared to nonhyperuniform structures, thus their permeabilities are generally lower[40].

Microstructural descriptors

We study the performance of the different microstructural descriptors introduced above and the combinations of them for predicting the permeability k (dimension length^2). The descriptors used are porosity ϕ (dimensionless), specific surface s (dimension 1/length), tortuosity τ (dimensionless), the correlation functions F_{ss} (dimension 1/length^2), F_{sv} (dimension 1/length), and F_{vv} (dimensionless). Additionally, we investigate a particular combination of the correlation functions: inspired by a rigorous upper bound for permeability in isotropic media[56], i.e.,

$$k \leq \frac{2}{3}\int_0^\infty \left[\frac{\phi^2}{s^2}F_{ss}(r) - \frac{2\phi}{s}F_{sv}(r) + F_{vv}(r)\right] r\mathrm{d}r, \tag{16}$$

we define a function

$$F(r) = \frac{\phi^2}{s^2}F_{ss}(r) - \frac{2\phi}{s}F_{sv}(r) + F_{vv}(r), \tag{17}$$

which is also used for prediction (F is dimensionless and converges to zero when r goes to infinity). Each correlation function is represented by a 96-dimensional vector. In Table 1, the models, denoted 1 through 7, and the sizes of the corresponding input features are listed. Additionally, we consider a rescaling of the problem, predicting the rescaled, dimensionless permeability ks^2 instead of k directly. For model 1, we remove s from the descriptors since it is already absorbed in the permeability; for models 2, 3, 4, 6, and 7, the correlation functions

No.	Descriptors	Input size
$1'$	ϕ, τ	2
$2'$	F_{ss}/s^2	96
$3'$	F_{sv}/s	96
$4'$	F_{vv}	96
$5'$	F	96
$6'$	$F_{ss}/s^2, F_{sv}/s, F_{vv}$	288
$7'$	$F_{ss}/s^2, F_{sv}/s, F_{vv}, \tau$	289

Table 2. Descriptors for prediction of the rescaled, dimensionless permeability ks^2.

are rescaled to dimensionless versions where applicable. In Table 2, the modified models, denoted $1'$ through $7'$, and the dimensions of the corresponding input vectors (that changes only for model $1'$) are listed.

In practice, we use the logarithm of permeabilities, i.e., $\log_{10} k$ and $\log_{10}\left(ks^2\right)$, instead of their original values. The reason for using logarithms of the permeabilities is that they span several orders of magnitude. By taking logarithms, the predictions are simplified, and guaranteed to be positive. We also use the logarithms of porosity, specific surface, and tortuosity since we know that models 1 and $1'$ are naturally multiplicative in these Kozeny-Carman-like equations.

Predictive models

We assess the predictive performance of the different descriptors/inputs using several regression methods, namely, linear regression with linear terms only or combined with quadratic terms, and deep artificial neural networks. The inputs are as described above, with no normalization (such as subtracting feature-wise means; our investigation suggested no improvement from normalization in this setting). For each microstructure class, the data are split randomly into training data (70 %; 7,000 per class), validation data (15 %; 1,500 per class), and test data (15 %; 1,500 per class). In total, the training, validation, and test data sets hence consist of 21,000, 4,500, and 4,500 samples, respectively. The split is kept fixed across all inputs and all regression methods. Training data is used for the actual estimation of a functional relationship mapping input to output. Validation data is used for hyperparameter selection, i.e., finding optimal values for e.g. learning rates for ANNs (in the case of linear regressions, the validation data is not used because we do not have any hyperparameters to optimize). Test data is used for final assessment of the predictive performance. To quantify error/loss in prediction, we use several different measures. Let k be the 'true' permeability, i.e., the value obtained from the lattice Boltzmann simulations, and let $\hat{k}$ be the predicted value. We use mean squared error (MSE) in the logarithmic scale,

$$\mathrm{MSE} = \left\langle \left(\log_{10}\hat{k} - \log_{10} k\right)^2 \right\rangle, \tag{18}$$

root mean squared error (RMSE) which is just $\mathrm{RMSE} = \mathrm{MSE}^{1/2}$, and mean absolute percentage error (MAPE) in the linear scale, i.e.,

$$\mathrm{MAPE} = 100 \times \left\langle \left| \frac{\hat{k} - k}{k} \right| \right\rangle \%. \tag{19}$$

Using MSE is the most practical and most common choice for model fitting. However, for final assessment of performance, the linear scale and MAPE is a more straightforward and intuitive choice.

Linear regression with linear terms. First, we consider using linear regression with only linear terms (i.e., only the input descriptors to the power of unity are used). For models 1 and $1'$, this becomes multiplicative regression in a Kozeny-Carman-like form, i.e.,

$$\log_{10} k = c_0 + a \log_{10}\phi + b \log_{10} s + c \log_{10}\tau \tag{20}$$

and

$$\log_{10}\left(ks^2\right) = c_0 + a \log_{10}\phi + c \log_{10}\tau. \tag{21}$$

It is well established that $a > 0$, $b < 0$, and $c < 0$ in this setting (and due to the dimensions, $b = -2$ would be preferable).

On the other hand, the rationale behind the linear regression model of correlation functions is inspired by the rigorous bounds involving correlation functions, such as Eq. (16). Since the integral in the bounds can be seen as the inner product between the correlation functions and another predetermined function, it is natural to assume that a functional regression on the correlation functions may yield a reasonable estimation of permeabilities. For correlation functions evaluated on discrete grids, the model essentially becomes a linear regression model. To give a couple of examples of the regressions on correlation functions, model 2 becomes

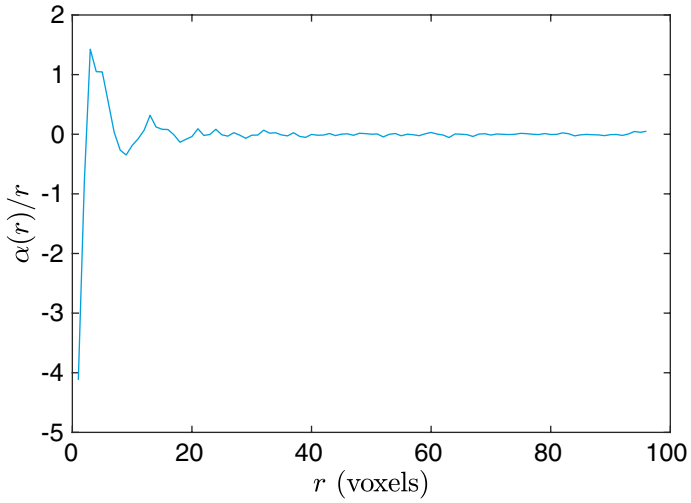

Figure 5. The estimated coefficient function $\alpha(r)$ scaled by r for model 4 using the linear regression model. One can see that it gives larger weight to the small-to-intermediate-r behavior of F_{vv}.

$$\log_{10} k = c_0 + \sum_i \alpha(r_i) F_{ss}(r_i), \tag{22}$$

model 6 becomes

$$\log_{10} k = c_0 + \sum_i \alpha(r_i) F_{ss}(r_i) + \sum_i \beta(r_i) F_{sv}(r_i) + \sum_i \gamma(r_i) F_{vv}(r_i), \tag{23}$$

and model 3′ becomes

$$\log_{10}\left(ks^2\right) = c_0 + \sum_i \alpha(r_i) F_{sv}(r_i)/s. \tag{24}$$

The rest of the models are formulated in an equivalent fashion. For the correlation function-based models, α, β, and γ are just vectors of coefficients but can also be thought of as discretized forms of continuous coefficient functions $\alpha(r)$, $\beta(r)$, and $\gamma(r)$. We use least squares fitting, finding the coefficients that minimize the training set MSE. We also include the reference Kozeny-Carman model in this category as a benchmark. Fitting is performed in Matlab (Mathworks, Natick, MA, US).

Specifically, the fitted Kozeny-Carman model writes as

$$k = \frac{\phi^3}{6.23 s^2}. \tag{25}$$

For models 1 and 1′, the estimated relationships become

$$k = 0.29 \frac{\phi^{2.68}}{s^{1.88} \tau^{7.28}}, \tag{26}$$

and

$$ks^2 = 0.21 \frac{\phi^{2.77}}{\tau^{6.36}}. \tag{27}$$

Interestingly, the estimated coefficients for models 1 and 1′ are quite similar, and they are also comparable to the Kozeny-Carman model. Specifically, even without being forced to have the dimensionally correct exponent − 2 for s, as in the case of model 1′, the estimated exponent from the regression (− 1.88) in model 1 is very close.

To get an idea of the dependence of the coefficient functions on r, we plot the estimated $\alpha(r)/r$ in Fig. 5 as an example. We scale $\alpha(r)$ by r in order to make it have the appropriate weight consistent with the one that multiplies the correlation functions in the integrand of Eq. (16). It is clear that instead of weighting on different parts of $F_{vv}(r)$ equally, the regression emphasizes on the small-to-intermediate-r behavior of the correlation function as expected.

We also made an attempt to further smooth the coefficient functions according to a proposed functional regression scheme published earlier[68]. Briefly, the idea is to introduce a penalty term in the least squares objective function that is proportional to the integral of the squared second derivative of the coefficient function, such that the functional form becomes smoother. To exemplify, pick a model with a coefficient function $\alpha(r)$, and we minimize

No.	RMSE Train	Val	Test	MAPE (%) Test
Kozeny-Carman model				
–	0.084	0.084	0.086	14.986
Unscaled models				
1	0.046	0.046	0.047	8.291
2	0.370	0.371	0.376	98.028
3	0.163	0.163	0.167	29.719
4	0.068	0.069	0.069	12.011
5	0.233	0.238	0.243	51.563
6	0.049	0.050	0.051	8.351
7	0.033	0.034	0.034	5.536
Rescaled models				
$1'$	0.052	0.051	0.052	8.653
$2'$	0.360	0.361	0.365	96.907
$3'$	0.078	0.078	0.081	14.363
$4'$	0.059	0.058	0.060	10.315
$5'$	0.165	0.164	0.168	33.823
$6'$	0.043	0.044	0.045	7.488
$7'$	0.029	0.030	0.031	5.063

Table 3. RMSE for the training, validation, and test sets and MAPE for the test set for the regression models with linear terms. We also include the reference Kozeny-Carman model in this category.

$$\sum_j \left(\log_{10}\hat{k}(j) - \log_{10}k(j)\right)^2 + \lambda \sum_i (\alpha(r_{i-1}) - 2\alpha(r_i) + \alpha(r_{i+1}))^2 \tag{28}$$

for a range of penalty parameters λ, and pick λ such that the validation MSE is minimized. Interestingly, the smoothing procedure provides only a negligible improvement in validation and test errors, and a negligible increase in smoothness of the coefficient functions. Thus, we stick to the models without penalties.

The errors for training, validation, and test sets for the aforementioned models are shown in Table 3. Although we perform no hyperparameter optimization for these regression models (such as variable selection), we include the validation set errors for consistency. There are several noteworthy observations about our results. First, the Kozeny-Carman-like model (1 and $1'$) with geodesic tortuosity almost reduce the relative error by half compared to the original Kozeny-Carman model, while the combination of all three correlation functions (6 and $6'$) also give comparable performance. Interestingly, when we further combine the correlation functions with geodesic tortuosity (7 and $7'$), the error shrinks to approximately only one-third of the error generated by the Kozeny-Carman model. Among single correlation functions, the best performance is given by F_{vv}, while the worst is given by F_{ss}. This result is expected, since F_{vv} alone can yield a bound on the permeability, while F_{ss} alone does not even contain the most important information, i.e., the porosity. The reason we keep model 2 and $2'$ is indeed just for self-consistency. Interestingly, the compound correlation function F performs relatively poorly. This is probably due to the fact that it washes out the information content contained in individual correlation functions. However, its error can be interpreted as a lower bound on how good the bound in Eq. (16) would work on our data set. To better visualize our findings, the predicted values vs the true (simulated) values and histograms of relative errors for a few selected models are shown in Fig. 6. It is obvious that by adding geodesic tortuosity and correlation functions the prediction error can be greatly reduced. It can also be seen that although the Kozeny-Carman-like model (Fig. 6b) and the correlation function based one (Fig. 6c) has similar MAPE, the error distribution of the correlation function based one is more symmetric.

Linear regression with quadratic terms. Second, we generalize the previous models that utilized linear terms only to incorporate both linear and quadratic terms. For the correlation function models, we use both the correlation functions themselves and their squares. For example, we use both F_{ss} and F_{ss}^2 as input data in model 2. It is worth to point out that we use only pure quadratic terms, such as $F_{ss}^2(r_i)$, but not mixed quadratic terms, such as $F_{ss}(r_i)F_{ss}(r_j)$ for $i \neq j$. We include models 1 and $1'$ in this investigation as well, mainly for completeness, and add terms of the type $\left(\log_{10}\phi\right)^2$. Fitting is performed in Matlab (Mathworks, Natick, MA, US). The errors for training, validation, and test sets are again shown in Table 4. In Fig. 7, the predicted values vs the true (simulated) values and histograms of relative errors for a few selected models are shown.

Importantly, we note that adding the quadratic terms leads to an improvement for every model. This suggests that the relation between the microstructural descriptors and the permeability can be quite complex such that a simple linear model may not be able to fully capture it. However, the relative rank of performances roughly remain the same, showing the validity of our previous arguments. The estimated coefficient functions are quite

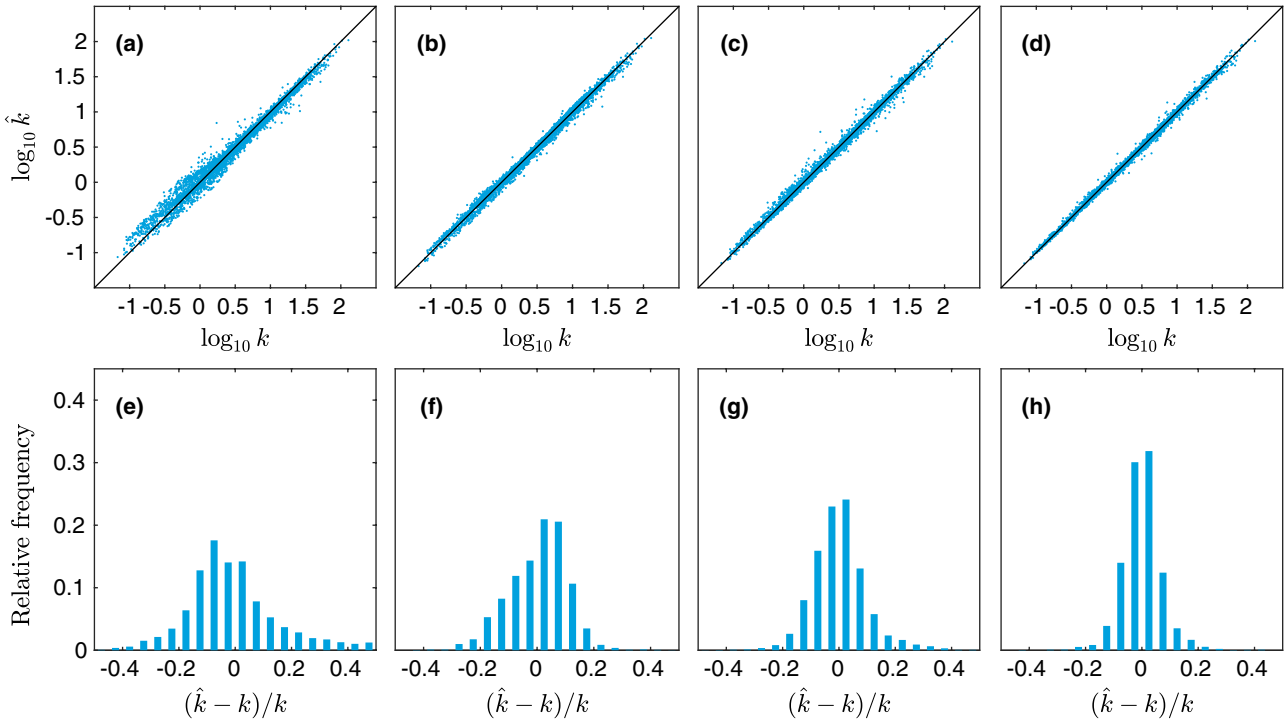

Figure 6. Predicted values $\log_{10}\hat{k}$ vs the true (simulated) values $\log_{10} k$ for linear regression with linear terms, showing (**a**) the Kozeny-Carman model, (**b**) model 1, (**c**) model $6'$, and (**d**) model $7'$. In (**e**–**h**), histograms of relative errors are shown for the same set of models.

No.	RMSE Train	Val	Test	MAPE (%) Test
Unscaled models				
1	0.038	0.038	0.039	6.560
2	0.351	0.351	0.359	92.349
3	0.079	0.079	0.081	13.547
4	0.059	0.061	0.061	10.624
5	0.151	0.156	0.163	29.633
6	0.039	0.040	0.040	6.377
7	0.027	0.027	0.028	4.300
Rescaled models				
$1'$	0.050	0.049	0.050	8.216
$2'$	0.358	0.360	0.365	96.739
$3'$	0.051	0.051	0.052	8.717
$4'$	0.051	0.052	0.053	8.923
$5'$	0.118	0.118	0.122	22.362
$6'$	0.036	0.039	0.039	6.041
$7'$	0.026	0.028	0.029	4.515

Table 4. RMSE for the training, validation, and test sets and MAPE for the test set for the regression models with linear and quadratic terms.

noisy and their physical meaning is not obvious. We also make an attempt to use a full quadratic model, incorporating also mixed terms such as $F_{ss}(r_i)F_{ss}(r_j)$, or even mixed between correlation functions, such as $F_{sv}(r_i)F_{vv}(r_j)$. The numbers of variables in the models then become very large, leading to ill-conditioned estimation problems. We investigated whether the Lasso variable selection technique[69], which forces a variable number of coefficients in a linear model to become zero by penalizing the sum of absolute values of the coefficients, could act as an efficient means of reducing the model dimensionality. However, it turns out that no amount of Lasso regularization can decrease the validation MSE in this case. There are two likely reasons for this: Lasso is primarily intended for

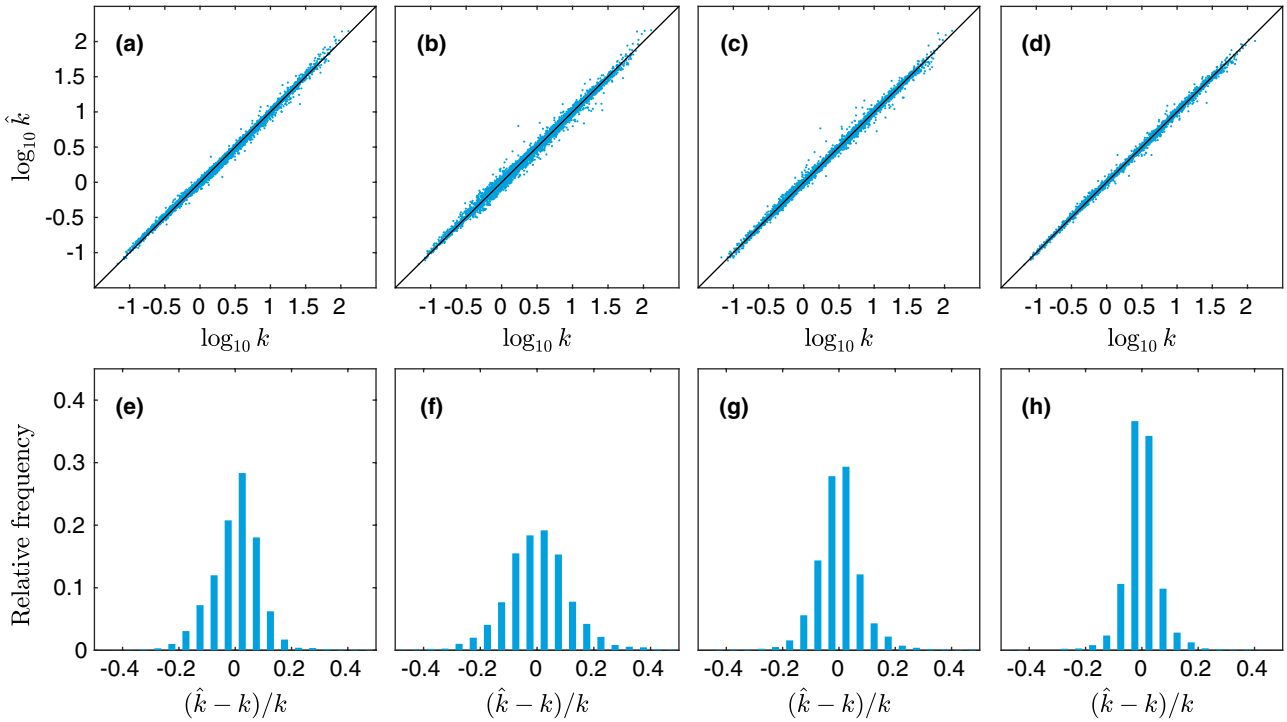

Figure 7. Predicted values $\log_{10}\hat{k}$ vs the true (simulated) values $\log_{10} k$ for linear regression with linear and quadratic terms, showing (**a**) model 1, (**b**) model 4′, (**c**) model 6′, and (**d**) model 7. In (**e**–**h**), histograms of relative errors are shown for the same set of models.

high-dimensional variable spaces where a large fraction of the variables contain little information and mostly noise, and can be disregarded easily. This is likely not the case for the correlation functions. Also, because the values are taken from continuous functions, they are strongly correlated, which is known to compromise the underlying rationale of Lasso.

Neural networks. The complexity of the linear regression models could be further increased, for example by incorporating pure cubic terms. Although we expect to see further improvements, the linear regression model can quickly become ill-conditioned and intractable on this track. For this reason, we proceed to consider deep neural networks, which can potentially fully capture the complex structure–property relationships. Thus we can exploit the complete information content contained in the descriptors.

We use four fully-connected hidden layers with 128 nodes each and rectified linear unit (ReLU) activations. Given that the input dimension is n, and the output dimension is unity, the number of weights in the network is $(n+1)\times 128 + 3\times 129\times 128 + 128 + 1$, i.e., there are between 50,049 and 86,785 weights to be optimized. The network is shown in Fig. 8. Random initial weights are selected using the Glorot/Xavier uniform initializer. The networks are trained using the Adam optimizer[70] with learning rate 10^{-4}, batch size 128, and mean squared error loss. All models are trained 100 times for 10,000 epochs using different random weight initializations, and the models with the globally minimal validation loss (MSE) are selected (hence utilizing early stopping, but performed over multiple realizations/initializations). The reason for this procedure is to minimize the impact of the random weight initializations. The models are implemented in TensorFlow 2.1.0 (http://www.tensorflow.org)[71]. An example of training and validation loss curves is shown in Fig. 9. Again, the errors for training, validation, and test sets are shown in Table 5. In Fig. 10, the predicted values vs the true (simulated) values are shown. Indeed, the neural networks based regressions perform noticeably better than the linear counterpart. Again, we see that the combination of correlation functions and geodesic tortuosity gives the best performance, achieving an impressive MAPE that is less than 4%. We also notice that all correlation function based models (except for F_{ss} and F for aforementioned reasons) perform very well, and all better than the Kozeny-Carman-like model.

To gain some understanding of the neural network and how the prediction is performed, we perform for the case of model 4 an analysis of the network's sensitivity with respect to perturbations in the input. Specifically, for the test set, we add random Gaussian noise to $F_{vv}(r)$ for one r value at a time. The perturbation in the output is quantified by the standard deviation of the difference between the original prediction and the perturbed prediction. In Fig. 11, we show the results for $\sigma = 0.02$ (it turns out that for a broad range of perturbations, $0.0001 \le \sigma \le 0.1$, the result changes only by a constant scaling). We see a rough resemblance to Fig. 5 in the sense that large magnitudes are mostly found for small r, indicating that to some extent the models are using the same information in the correlation function.

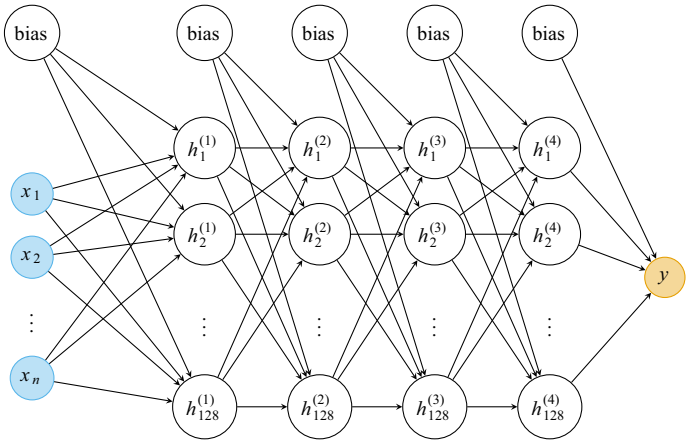

Figure 8. The topology of the neural network. The input variables are denoted x_1 to x_n, where the input dimension n is either 2, 3, 96, 288, or 289. There are four fully-connected hidden layers with 128 nodes each. The k:th node of the l:th layer is here denoted by h_k^l. Rectified linear unit (ReLU) activations are used for the hidden layers. The output y is just the logarithm of the permeability. The figure is produced by Victor Wåhlstrand Skärström in TikZ/LaTeX (MikTeX distribution version 20.6.29, http://www.miktex.org, freely available without permission).

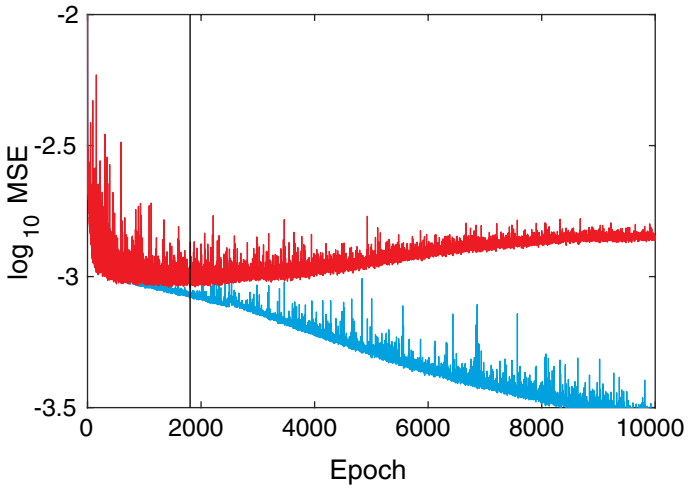

Figure 9. An example of a training (blue) and validation (red) loss curves for model 6′. This is the best run for this model, i.e., the one that yielded the minimum validation MSE, which is indicated by the vertical line (black).

Conclusions and discussion

We have studied data-driven structure–property relationships between fluid permeabilities and a variety of microstructural descriptors in a large data set of 30,000 virtual, porous microstructures of different types. The data set includes both granular and continuous solid phases, and is the largest one ever generated for the study of transport properties to our knowledge. To characterize the pore space geometry, we computed one-point correlation functions (porosity, specific surface), two-point surface-surface, surface-void, and void-void correlation functions, and geodesic tortuosity. Different combinations of these descriptors were used as input for different statistical learning methods, including linear regression with linear and quadratic terms, as well as deep neural networks. We find that the performance improves as the regression models become more complex, suggesting the complex relationship between the structural descriptors and the physical properties. Sufficiently large neural networks are able to fully capture the information content of the descriptors and reveal their utilities. With higher-order descriptors, we obtain significant improvements of performance when compared to a Kozeny-Carman regression with only lowest-order descriptors (porosity and specific surface). We found that combining all three two-point correlation functions and tortuosity provides the best prediction of permeability. The void-void correlation function was found to be the most informative individual descriptor. Also, the combination of porosity, specific surface, and geodesic tortuosity provides comparable predictive performance, in spite of its simplicity. Indeed, this shows that the greater information content contained in higher-order correlation functions are

No.	RMSE Train	Val	Test	MAPE (%) Test
Unscaled models				
1	0.029	0.030	0.041	6.374
2	0.317	0.322	0.331	76.975
3	0.032	0.035	0.037	5.479
4	0.031	0.032	0.033	4.705
5	0.057	0.066	0.066	10.760
6	0.028	0.030	0.032	4.431
7	0.021	0.023	0.025	3.679
Rescaled models				
$1'$	0.045	0.045	0.046	7.857
$2'$	0.326	0.330	0.339	82.826
$3'$	0.032	0.037	0.039	6.101
$4'$	0.031	0.034	0.036	5.179
$5'$	0.056	0.067	0.069	11.080
$6'$	0.029	0.032	0.036	5.176
$7'$	0.020	0.026	0.028	4.133

Table 5. RMSE for the training, validation, and test sets and MAPE for the test set for the ANN models.

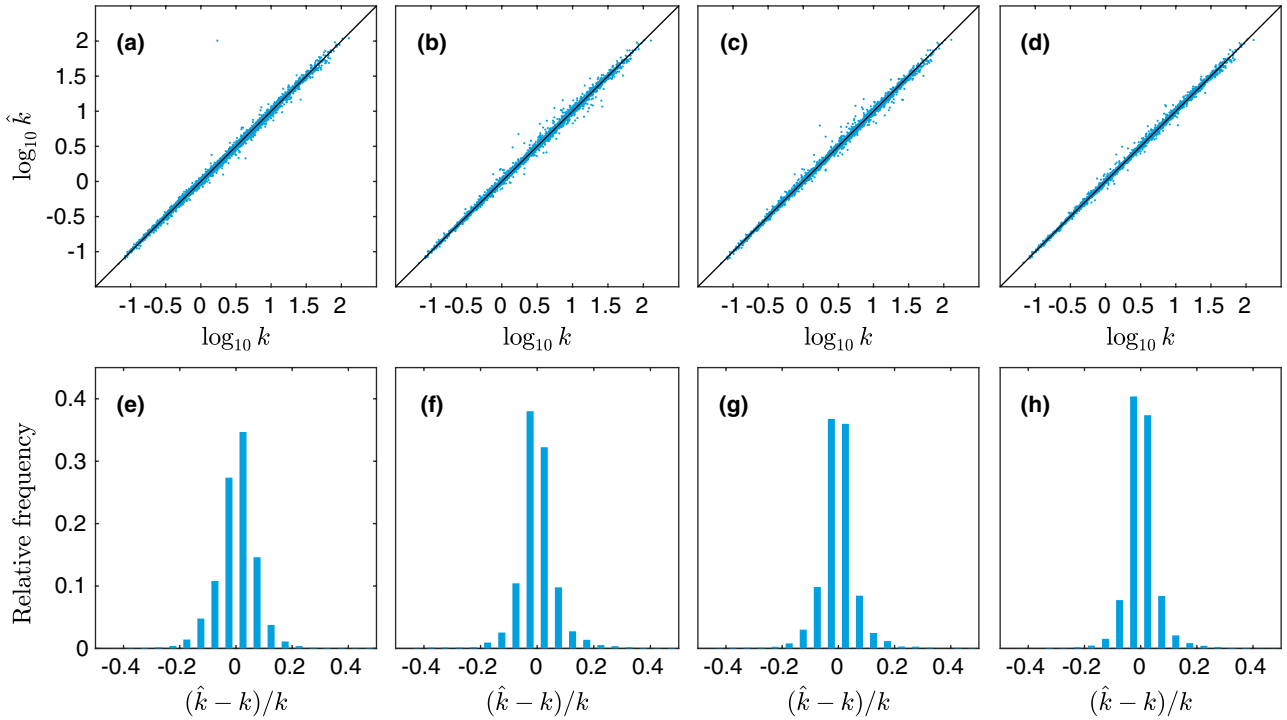

Figure 10. Predicted values $\log_{10}\hat{k}$ vs the true (simulated) values $\log_{10} k$ for deep neural network regression, showing (**a**) model 1, (**b**) model 4, (**c**) model 6, and (**d**) model 7. In (**e**–**h**), histograms of relative errors are shown for the same set of models.

extremely useful for predicting physical properties of complex materials. Moreover, our work demonstrates that advanced machine learning methods can be very useful in establishing structure–property relationships.

An interesting observation is that in general the rescaling of permeabilities seem to improve the performance of the simple linear model, as seen from Table 3 that all models except the Kozeny-Carman-like one outperform their unscaled counterparts. However, as we add quadratic terms and the predictive model becomes more complex, the advantage of rescaling does not hold any more. Finally, for the highly nonlinear neural networks, the relative performance is completely inverted. This may suggest that the relation between the permeability and the specific surface cannot be captured by a simple rescaling, and doing so may reduce the information content

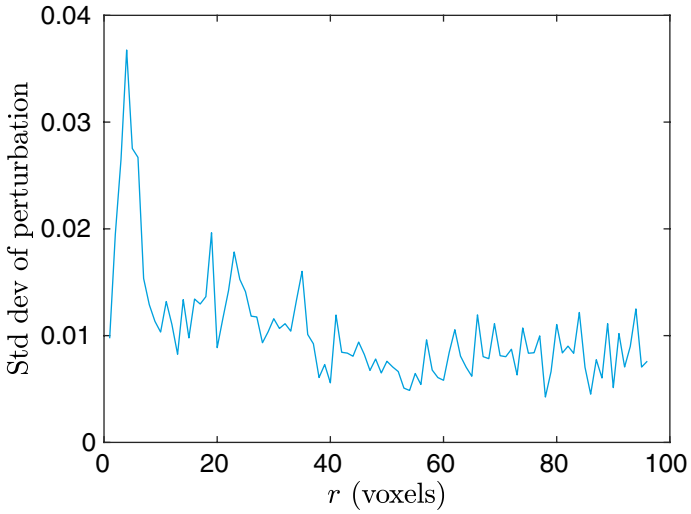

Figure 11. The standard deviation of the perturbation in the output of the neural network for model 4, as a function of the r value where the perturbation in the input is introduced.

contained in the original data. These observed effects of rescaling may lead to some guidelines on developing physics-aware machine learning models for other physical properties as well.

As a final remark, we emphasize that by incorporating microstructures with different length scales in our data set we make our models very robust and can be applied to real-world data. Since the permeability has the dimension L^2, we can easily obtain the permeability of a sample with a different length scale but the same microstructure. However, there is no universally applicable characteristic microstructural length to scale the permeability that enables a comparison of permeabilities for different microstructures. For example, for sphere packings the natural choice can be the radii of particles, but for continuous structures we need to resort to other quantities[40]. Thus, by training our models on samples with varying length scales, we circumvent this problem by only requiring a rescaling to the right order of magnitude. Finally, we make the data and code used publicly available to facilitate further development of permeability prediction methods[54].

Data availability

The data, i.e. microstructural descriptors and computed permeabilities, together with the trained models and the code used are publicly available via a repository[54].

Received: 1 June 2020; Accepted: 25 August 2020
Published online: 17 September 2020

References

1. Torquato, S. *Random Heterogeneous Materials: Microstructure and Macroscopic Properties* (Springer, New York, 2013).
2. Vasseur, J., Wadsworth, F. B. & Dingwell, D. B. Permeability of polydisperse magma foam. *Geology* **48**(6), 536–540 (2020).
3. Silvestre, C., Duraccio, D. & Cimmino, S. Food packaging based on polymer nanomaterials. *Prog. Polym. Sci.* **36**, 1766–1782 (2011).
4. Slater, A. & Cooper, A. Function-led design of new porous materials. *Science* **348**, aaa8075 (2015).
5. Stamenkovic, V., Strmcnik, D., Lopes, P. & Markovic, N. Energy and fuels from electrochemical interfaces. *Nat. Mater.* **16**, 57–69 (2017).
6. van Langenhove, L. *Smart Textiles for Medicine and Healthcare: Materials, Systems and Applications* (Elsevier, Amsterdam, 2007).
7. Marucci, M. *et al.* New insights on how to adjust the release profile from coated pellets by varying the molecular weight of ethyl cellulose in the coating film. *Int. J. Pharm.* **458**, 218–223 (2013).
8. Milton, G. & Sawicki, A. Theory of composites. Cambridge monographs on applied and computational mathematics. *Appl. Mech. Rev.* **56**, B27–B28 (2003).
9. Sahimi, M. *Flow and Transport in Porous Media and Fractured Rock: From Classical Methods to Modern Approaches* (Wiley, Hoboken, 2011).
10. Huang, S., Wu, Y., Meng, X., Liu, L. & Ji, M. Recent advances on microscopic pore characteristics of low permeability sandstone reservoirs. *Adv. Geo-Energy Res.* **2**, 122–134 (2018).
11. Huang, H. *et al.* Effects of pore-throat structure on gas permeability in the tight sandstone reservoirs of the Upper Triassic Yanchang formation in the Western Ordos Basin. *China. J. Petrol. Sci. Eng.* **162**, 602–616 (2018).
12. Blunt, M. J. *et al.* Pore-scale imaging and modelling. *Adv. Water Resour.* **51**, 197–216 (2013).
13. Lee, S.-H., Chang, W.-S., Han, S.-M., Kim, D.-H. & Kim, J.-K. Synchrotron x-ray nanotomography and three-dimensional nanoscale imaging analysis of pore structure-function in nanoporous polymeric membranes. *J. Membr. Sci.* **535**, 28–34 (2017).
14. Gunda, N. *et al.* Focused ion beam-scanning electron microscopy on solid-oxide fuel-cell electrode: Image analysis and computing effective transport properties. *J. Power Sources* **196**, 3592–3603 (2011).
15. Ge, X., Fan, Y., Zhu, X., Chen, Y. & Li, R. Determination of nuclear magnetic resonance T2 cutoff value based on multifractal theory—An application in sandstone with complex pore structure. *Geophysics* **80**, D11–D21 (2015).
16. Yao, Y. & Liu, D. Comparison of low-field NMR and mercury intrusion porosimetry in characterizing pore size distributions of coals. *Fuel* **95**, 152–158 (2012).

17. Kozeny, J. . Über kapillare leitung des wassers im boden:(aufstieg, versickerung und anwendung auf die bewässerung). *Sitz. Ber. Akad. Wiss, Wien, Math. Nat.* **136**, 271–306 (1927).
18. Carman, P. Fluid flow through granular beds. *Trans. Inst. Chem. Eng.* **15**, 150–166 (1937).
19. Kaviany, M. *Principles of heat transfer in porous media* (Springer, New York, 2012).
20. Xu, P. & Yu, B. Developing a new form of permeability and Kozeny-Carman constant for homogeneous porous media by means of fractal geometry. *Adv. Water Resour.* **31**, 74–81 (2008).
21. Mauret, E. . & Renaud, M. . Transport phenomena in multi-particle systems—I. Limits of applicability of capillary model in high voidage beds-application to fixed beds of fibers and fluidized beds of spheres. *Chem. Eng. Sci.* **52**, 1807–1817 (1997).
22. Mota, M., Teixeira, J., Bowen, W. & Yelshin, A. Binary spherical particle mixed beds: Porosity and permeability relationship measurement. *Trans. Filtr. Soc.* **1**, 101–106 (2001).
23. Plessis, J. D. & Masliyah, J. Flow through isotropic granular porous media. *Transport Porous Med.* **6**, 207–221 (1991).
24. Ahmadi, M., Mohammadi, S. & Hayati, A. Analytical derivation of tortuosity and permeability of monosized spheres: A volume averaging approach. *Phys. Rev. E* **83**, 026312 (2011).
25. Jiao, Y., Stillinger, F. & Torquato, S. A superior descriptor of random textures and its predictive capacity. *Proc. Natl. Acad. Sci.* **106**, 17634–17639 (2009).
26. Gommes, C., Jiao, Y. & Torquato, S. Microstructural degeneracy associated with a two-point correlation function and its information content. *Phys. Rev. E* **85**, 051140 (2012).
27. Torquato, S. Random heterogeneous media: Microstructure and improved bounds on effective properties. *Appl. Mech. Rev.* **44**, 37–76 (1991).
28. Jiao, Y. & Torquato, S. Quantitative characterization of the microstructure and transport properties of biopolymer networks. *Phys. Biol.* **9**, 036009 (2012).
29. Prager, S. Viscous flow through porous media. *Phys. Fluids* **4**, 1477–1482 (1961).
30. Weissberg, H. & Prager, S. Viscous flow through porous media. II. approximate three-point correlation function. *Phys. Fluids* **5**, 1390–1392 (1962).
31. Weissberg, H. & Prager, S. Viscous flow through porous media. III. Upper bounds on the permeability for a simple random geometry. *Phys. Fluids* **13**, 2958–2965 (1970).
32. Berryman, J. & Milton, G. Normalization constraint for variational bounds on fluid permeability. *J. Chem. Phys.* **83**, 754–760 (1985).
33. Berryman, J. Bounds on fluid permeability for viscous flow through porous media. *J. Chem. Phys.* **82**, 1459–1467 (1985).
34. Rubinstein, J. & Torquato, S. Flow in random porous media: Mathematical formulation, variational principles, and rigorous bounds. *J. Fluid Mech.* **206**, 25–46 (1989).
35. Liasneuski, H. *et al.* Impact of microstructure on the effective diffusivity in random packings of hard spheres. *J. Appl. Phys.* **116**, 034904 (2014).
36. Hlushkou, D., Liasneuski, H., Tallarek, U. & Torquato, S. Effective diffusion coefficients in random packings of polydisperse hard spheres from two-point and three-point correlation functions. *J. Appl. Phys.* **118**, 124901 (2015).
37. Zachary, C. & Torquato, S. Improved reconstructions of random media using dilation and erosion processes. *Phys. Rev. E* **84**, 056102 (2011).
38. Guo, E.-Y., Chawla, N., Jing, T., Torquato, S. & Jiao, Y. Accurate modeling and reconstruction of three-dimensional percolating filamentary microstructures from two-dimensional micrographs via dilation-erosion method. *Mater. Charact.* **89**, 33–42 (2014).
39. Katz, A. & Thompson, A. Quantitative prediction of permeability in porous rock. *Phys. Rev. B* **34**, 8179 (1986).
40. Torquato, S. Predicting transport characteristics of hyperuniform porous media via rigorous microstructure-property relations. *Adv. Water Resour.* **140**, 103565 (2020).
41. Avellaneda, M. & Torquato, S. Rigorous link between fluid permeability, electrical conductivity, and relaxation times for transport in porous media. *Phys. Fluids A Fluid Dyn.* **3**, 2529–2540 (1991).
42. Ghanbarian, B., Hunt, A., Ewing, R. & Sahimi, M. Tortuosity in porous media: A critical review. *Soil Sci. Soc. Am. J.* **77**, 1461–1477 (2013).
43. van der Linden, J., Narsilio, G. & Tordesillas, A. Machine learning framework for analysis of transport through complex networks in porous, granular media: A focus on permeability. *Phys. Rev. E* **94**, 022904 (2016).
44. Stenzel, O., Pecho, O., Holzer, L., Neumann, M. & Schmidt, V. Predicting effective conductivities based on geometric microstructure characteristics. *AIChE J.* **62**, 1834–1843 (2016).
45. Neumann, M., Stenzel, O., Willot, F., Holzer, L. & Schmidt, V. Quantifying the influence of microstructure on effective conductivity and permeability: virtual materials testing. *Int. J. Solids Struct.* (2019).
46. Barman, S., Rootzén, H. & Bolin, D. Prediction of diffusive transport through polymer films from characteristics of the pore geometry. *AIChE J.* **65**, 446–457 (2019).
47. Kondo, R., Yamakawa, S., Masuoka, Y., Tajima, S. & Asahi, R. Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics. *Acta Mater.* **141**, 29–38 (2017).
48. Wu, J., Yin, X. & Xiao, H. Seeing permeability from images: Fast prediction with convolutional neural networks. *Sci. Bull.* **63**, 1215–1222 (2018).
49. Sudakov, O., Burnaev, E. & Koroteev, D. Driving digital rock towards machine learning: Predicting permeability with gradient boosting and deep neural networks. *Comput. Geosci.* **127**, 91–98 (2019).
50. Stenzel, O., Pecho, O., Holzer, L., Neumann, M. & Schmidt, V. Big data for microstructure-property relationships: A case study of predicting effective conductivities. *AIChE J.* **63**, 4224–4232 (2017).
51. Kamrava, S., Tahmasebi, P. & Sahimi, M. Linking morphology of porous media to their macroscopic permeability by deep learning. *Transport Porous Med.* **131**, 427–448 (2020).
52. Wu, H., Fang, W.-Z., Kang, Q., Tao, W.-Q. & Qiao, R. Predicting effective diffusivity of porous media from images by deep learning. *Sci. Rep.* **9**, 20387 (2019).
53. Lubbers, N., Lookman, T. & Barros, K. Inferring low-dimensional microstructure representations using convolutional neural networks. *Phys. Rev. E* **96**, 052111 (2017).
54. Röding, M., Ma, Z. & Torquato, S. Predicting permeability via statistical learning on higher-order microstructural information. *ZENODO* https://doi.org/10.5281/zenodo.3752765 (2020).
55. Pecho, O. *et al.* 3D microstructure effects in Ni-YSZ anodes: Prediction of effective transport properties and optimization of redox stability. *Materials* **8**, 5554–5585 (2015).
56. Ma, Z. & Torquato, S. Precise algorithms to compute surface correlation functions of two-phase heterogeneous media and their applications. *Phys. Rev. E* **98**, 013307 (2018).
57. Scholz, C. *et al.* Direct relations between morphology and transport in Boolean models. *Phys. Rev. E* **92**, 043023 (2015).
58. Howard, M. *et al.* Connecting solute diffusion to morphology in triblock copolymer membranes. *Macromolecules* **53**(7), 2336–2343 (2020).
59. Lang, A. & Potthoff, J. Fast simulation of gaussian random fields. *Monte Carlo Methods Appl.* **17**, 195–214 (2011).
60. Matérn, B. *Spatial Variation* (Springer, New York, 1986).
61. Gebäck, T. & Heintz, A. A lattice Boltzmann method for the advection–diffusion equation with Neumann boundary conditions. *Commun. Comput. Phys.* **15**, 487–505 (2014).

62. Gebäck, T., Marucci, M., Boissier, C., Arnehed, J. & Heintz, A. Investigation of the effect of the tortuous pore structure on water diffusion through a polymer film using lattice Boltzmann simulations. *J. Phys. Chem. B* **119**, 5220–5227 (2015).
63. Perram, J. & Wertheim, M. Statistical mechanics of hard ellipsoids. I. Overlap algorithm and the contact function. *J. Comput. Phys.* **58**, 409–416 (1985).
64. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. Julia: A fresh approach to numerical computing. *SIAM Rev.* **59**, 65–98 (2017).
65. Ginzburg, I., Verhaeghe, F. & d'Humieres, D. Study of simple hydrodynamic solutions with the two-relaxation-times lattice Boltzmann scheme. *Commun. Comput. Phys.* **3**, 519–581 (2008).
66. Zou, Q. & He, X. On pressure and velocity boundary conditions for the lattice Boltzmann BGK model. *Phys. Fluids* **9**, 1591–1598 (1997).
67. Ma, Z. & Torquato, S. Random scalar fields and hyperuniformity. *J. Appl. Phys.* **121**, 244904 (2017).
68. Röding, M., Svensson, P. & Lorén, N. Functional regression-based fluid permeability prediction in monodisperse sphere packings from isotropic two-point correlation functions. *Comput. Mater. Sci.* **134**, 126–131 (2017).
69. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
70. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980 (2014).
71. Abadi, M. *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems (2015). https://www.tensorflow.org/. Software available from tensorflow.org.

Acknowledgements

M.R. acknowledges the financial support of the Swedish Research Council (Grant Number 2016-03809) and the Swedish Research Council for Sustainable Development (Grant Number 2019-01295). Z. M. and S. T. acknowledge the support of the Air Force Office of Scientific Research Program on Mechanics of Multifunctional Materials and Microsystems under Award No. FA9550-18-1-0514. The computations were in part performed on resources at Chalmers Centre for Computational Science and Engineering (C3SE) provided by the Swedish National Infrastructure for Computing (SNIC). A GPU used for part of this research was donated by the NVIDIA Corporation. Tobias Gebäck is acknowledged for help concerning the lattice Boltzmann computations. Victor Wåhlstrand Skärström is acknowledged for producing Fig. 8.

Author contributions

M.R., Z.M., and S.T. conceived the project and designed the study, and S.T. supervised the work. M.R. and Z.M. performed the computations and analysis. M.R., Z.M, and S.T. wrote and prepared the manuscript. M.R. and Z.M. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.